18.175 Theory of Probability
Fall 2008

Section 20

Kantorovich-Rubinstein Theorem.

Let (S,d) be a separable metric space. Denote by $\mathcal{P}_1(S)$ the set of all laws on S such that for some $z \in S$ (equivalently, for all $z \in S$),

$$\int_S d(x,z)\mathbb{P}(x) < \infty.$$

Let us denote by

$$M(\mathbb{P},\mathbb{Q}) = \Big\{\mu \,:\, \mu \text{ is a law on } S \times S \text{ with marginals } \mathbb{P} \text{ and } \mathbb{Q}\Big\}.$$

Definition. For $\mathbb{P},\mathbb{Q} \in \mathcal{P}_1(S)$, the quantity

$$W(\mathbb{P},\mathbb{Q}) = \inf\Big\{\int d(x,y)d\mu(x,y) : \mu \in M(\mathbb{P},\mathbb{Q})\Big\}$$

is called the *Wasserstein* distance between $\mathbb{P}$ and $\mathbb{Q}$.

A measure $\mu \in M(\mathbb{P},\mathbb{Q})$ represents a *transportation* between measures $\mathbb{P}$ and $\mathbb{Q}$. We can think of the conditional distribution $\mu(y|x)$ as a way to redistribute the mass in the neighborhood of a point x so that the distribution $\mathbb{P}$ will be redistributed to the distribution $\mathbb{Q}$. If the distance $d(x,y)$ represents the cost of moving x to y then the Wasserstein distance gives the optimal total cost of transporting $\mathbb{P}$ to $\mathbb{Q}$.

Given any two laws $\mathbb{P}$ and $\mathbb{Q}$ on S, let us define

$$\gamma(\mathbb{P},\mathbb{Q}) = \sup\Big\{\Big|\int f d\mathbb{P} - \int f d\mathbb{Q}\Big| : ||f||_{\mathrm{L}} \le 1\Big\}$$

and

$$m_d(\mathbb{P},\mathbb{Q}) = \sup\Big\{\int f d\mathbb{P} + \int g d\mathbb{Q} \,:\, f,g \in C(S),\ f(x)+g(y) < d(x,y)\Big\}.$$

Lemma 40 *We have* $\gamma(\mathbb{P},\mathbb{Q}) = m_d(\mathbb{P},\mathbb{Q})$.

Proof. Given a function f such that $||f||_{\mathrm{L}} \le 1$ let us take a small $\varepsilon > 0$ and $g(y) = -f(y) - \varepsilon$. Then

$$f(x) + g(y) = f(x) - f(y) - \varepsilon \le d(x,y) - \varepsilon < d(x,y)$$

and

$$\int f d\mathbb{P} + \int g d\mathbb{Q} = \int f d\mathbb{P} - \int f d\mathbb{Q} - \varepsilon.$$

Combining with the choice of $-f(x)$ and $g(y) = f(y) - \varepsilon$ we get

$$\Big|\int f d\mathbb{P} - \int f d\mathbb{Q}\Big| \le \sup\Big\{\int f d\mathbb{P} + \int g d\mathbb{Q} : f(x)+g(y) < d(x,y)\Big\} + \varepsilon$$

which, of course, proves that

$$\gamma(\mathbb{P},\mathbb{Q}) \le \sup\Bigl\{\int f d\mathbb{P} + \int g d\mathbb{Q} : f(x)+g(y) < d(x,y)\Bigr\}.$$

Let us now consider functions f, g such that $f(x)+g(y) < d(x,y)$. Define

$$e(x) = \inf_y (d(x,y) - g(y)) = -\sup_y (g(y) - d(x,y))$$

Clearly,

$$f(x) \le e(x) \le d(x,x) - g(x) = -g(x)$$

and, therefore,

$$\int f d\mathbb{P} + \int g d\mathbb{Q} \le \int e d\mathbb{P} - \int e d\mathbb{Q}.$$

Function e satisfies

$$\begin{aligned} e(x) - e(x') &= \sup_y \bigl(g(y) - d(x',y)\bigr) - \sup_y \bigl(g(y) - d(x,y)\bigr) \\ &\le \sup_y \bigl(d(x,y) - d(x',y)\bigr) \le d(x,x') \end{aligned}$$

which means that $||e||_{\mathrm{L}} = 1$. This finishes the proof.

□

We will need the following version of the Hahn-Banach theorem.

Theorem 48 *(Hahn-Banach) Let V be a normed vector space, E - a linear subspace of V and U - an open convex set in V such that $U \cap E \neq \emptyset$. If $r : E \to \mathbb{R}$ is a linear non-zero functional on E then there exists a linear functional $\rho : V \to \mathbb{R}$ such that $\rho|_E = r$ and $\sup_U \rho(x) = \sup_{U\cap E} r(x)$.*

Proof. Let $t = \sup\{r(x) : x \in U \cap E\}$ and let $B = \{x \in E : r(x) > t\}$. Since B is convex and $U \cap B = \emptyset$, the Hahn-Banach separation theorem implies that there exists a linear functional $q : V \to \mathbb{R}$ such that $\sup_U q(x) \le \inf_B q(x)$. For any $x_0 \in U \cap E$ let $F = \{x \in E : q(x) = q(x_0)\}$. Since $q(x_0) < \inf_B q(x)$, $F \cap B = \emptyset$. This means that the hyperplanes $\{x \in E : q(x) = q(x_0)\}$ and $\{x \in E : r(x) = t\}$ in the subspace E are parallel and this implies that $q(x) = \alpha r(x)$ on E for some $\alpha \neq 0$. Let $\rho = q/\alpha$. Then $r = \rho|_E$ and

$$\sup_U \rho(x) = \frac{1}{\alpha}\sup_U q(x) \le \frac{1}{\alpha}\inf_B q(x) = \inf_B r(x) = t = \sup_{U\cap E} r(x).$$

Since $r = \rho|_E$, this finishes the proof.

□

Theorem 49 *If S is a compact metric space then $W(\mathbb{P},\mathbb{Q}) = m_d(\mathbb{P},\mathbb{Q})$ for $\mathbb{P},\mathbb{Q} \in \mathcal{P}_1(S)$.*

Proof. Consider a vector space $V = C(S\times S)$ equipped with $\|\cdot\|_\infty$ norm and let

$$U = \bigl\{f \in V : f(x,y) < d(x,y)\bigr\}.$$

Obviously, U is convex and open because $S \times S$ is compact and any continuous function on a compact achieves its maximum. Consider a linear subspace E of V defined by

$$E = \bigl\{\phi \in V\ :\ \phi(x,y) = f(x) + g(y)\bigr\}$$

so that

$$U \cap E = \bigl\{f(x) + g(y) < d(x,y)\bigr\}.$$

Define a linear functional r on E by

$$r(\phi) = \int f d\mathbb{P} + \int g d\mathbb{Q} \quad \text{if} \quad \phi = f(x) + g(y).$$

By the above Hahn-Banach theorem, r can be extended to $\rho : V \to \mathbb{R}$ such that $\rho|_E = r$ and

$$\sup_U \rho(\phi) = \sup_{U \cap E} r(\phi) = m_d(\mathbb{P}, \mathbb{Q}).$$

Let us look at the properties of this functional. First of all, if $a(x,y) \geq 0$ then $\rho(a) \geq 0$. Indeed, for any $c \geq 0$

$$U \ni d(x,y) - c \cdot a(x,y) - \varepsilon < d(x,y)$$

and, therefore, for all $c \geq 0$

$$\rho(d - ca - \varepsilon) = \rho(d) - c\rho(a) - \rho(\varepsilon) \leq \sup_U \rho < \infty.$$

This can hold only if $\rho(a) \geq 0$. This implies that if $\phi_1 \leq \phi_2$ then $\rho(\phi_1) \leq \rho(\phi_2)$. For any function ϕ, both $-\phi, \phi \leq \|\phi\|_\infty \cdot 1$ and, by monotonicity of ρ,

$$|\rho(\phi)| \leq ||\phi||_\infty \rho(1) = ||\phi||_\infty.$$

Since $S \times S$ is compact and ρ is a continuous functional on $(C(S \times S), \| \cdot \|_\infty)$, by the Reisz representation theorem there exists a unique measure μ on the Borel σ-algebra on $S \times S$ such that

$$\rho(f) = \int f(x,y) d\mu(x,y).$$

Since $\rho|_E = r$,

$$\int (f(x) + g(y)) d\mu(x,y) = \int f d\mathbb{P} + \int g d\mathbb{Q}$$

which implies that $\mu \in M(\mathbb{P}, \mathbb{Q})$. We have

$$m_d(\mathbb{P}, \mathbb{Q}) = \sup_U \rho(\phi) = \sup \left\{ \int f(x,y) d\mu(x,y) : f(x,y) < d(x,y) \right\} = \int d(x,y) d\mu(x,y) \geq W(\mathbb{P}, \mathbb{Q}).$$

The opposite inequality is easy because for any f, g such that $f(x) + g(y) < d(x,y)$ and any $\nu \in M(\mathbb{P}, \mathbb{Q})$,

$$\int f d\mathbb{P} + \int g d\mathbb{Q} = \int (f(x) + g(y)) d\nu(x,y) \leq \int d(x,y) d\nu(x,y). \tag{20.0.1}$$

This finishes the proof and, moreover, it shows that the infimum in the definition of W is achieved on μ.

□

Remark. Notice that in the proof of this theorem we never used the fact that d is a metric. Theorem holds for any $d \in C(S \times S)$ under the corresponding integrability assumptions. For example, one can consider loss functions of the type $d(x,y)^p$ for $p > 1$, which are not necessarily metrics. However, in Lemma 40, the fact that d is a metric was essential.

Our next goal will be to show that $W = \gamma$ on separable and not necessarily compact metric spaces. We start with the following.

Lemma 41 *If (S, d) is a separable metric space then W and γ are metrics on $\mathcal{P}_1(S)$.*

Proof. Since for a bounded Lipschitz metric β we have $\beta(\mathbb{P}, \mathbb{Q}) \leq \gamma(\mathbb{P}, \mathbb{Q})$, γ is also a metric because if $\gamma(\mathbb{P}, \mathbb{Q}) = 0$ then $\beta(\mathbb{P}, \mathbb{Q}) = 0$ and, therefore, $\mathbb{P} = \mathbb{Q}$. As in (20.0.1), it should be obvious that $\gamma(\mathbb{P}, \mathbb{Q}) = m_d(\mathbb{P}, \mathbb{Q}) \leq W(\mathbb{P}, \mathbb{Q})$ and if $W(\mathbb{P}, \mathbb{Q}) = 0$ then $\gamma(\mathbb{P}, \mathbb{Q}) = 0$ and $\mathbb{P} = \mathbb{Q}$. Symmetry of W is obvious. It remains to show that $W(\mathbb{P}, \mathbb{Q})$ satisfies the triangle inequality. The idea will be rather simple, but to have well-defined

conditional distributions we will need to approximate distributions on $S \times S$ with given marginals by a more regular disributions with the same marginals. Let us first explain the main idea. Consider three laws $\mathbb{P}, \mathbb{Q}, \mathbb{T}$ on S and let $\mu \in M(\mathbb{P}, \mathbb{Q})$ and $\nu \in M(\mathbb{Q}, \mathbb{T})$ be such that

$$\int d(x,y)d\mu(x,y) \le W(\mathbb{P}, \mathbb{Q}) + \varepsilon \quad \text{and} \quad \int d(y,z)d\nu(y,z) \le W(\mathbb{Q}, \mathbb{T}) + \varepsilon.$$

Let us generate a distribution γ on $S \times S \times S$ with marginals $\mathbb{P}, \mathbb{Q}$ and $\mathbb{T}$ and marginals on pairs of coordinates (x, y) and (y, z) given by μ and ν by "gluing" μ and ν in the following way. Let us generate y from distribution $\mathbb{Q}$ and, given y, generate x and z according to conditional distributions $\mu(x|y)$ and $\nu(z|y)$ independently of each other, i.e.

$$\gamma(x, z|y) = \mu(x|y) \times \nu(z|y).$$

Obviously, by construction, (x, y) has distribution μ and (y, z) has distribution ν. Therefore, the marginals of x and z are $\mathbb{P}$ and $\mathbb{T}$ which means that the pair (x, z) has distribution $\eta \in M(\mathbb{P}, \mathbb{T})$. Finally,

$$\begin{aligned} W(\mathbb{P}, \mathbb{T}) &\le \int d(x,z)d\eta(x,z) = \int d(x,z)d\gamma(x,y,z) \le \int d(x,y)d\gamma + \int d(y,z)d\gamma \\ &= \int d(x,y)d\mu + \int d(y,z)d\nu \le W(\mathbb{P}, \mathbb{Q}) + W(\mathbb{Q}, \mathbb{T}) + 2\varepsilon. \end{aligned}$$

Letting $\varepsilon \to 0$ proves the triangle inequality for W. It remains to explain how the conditional distributions can be well defined. Let us modify μ by 'discretizing' it without losing much in the transportation cost integral. Given $\varepsilon > 0$, consider a partition $(S_n)_{n \ge 1}$ of S such that diameter$(S_n) < \varepsilon$ for all n. This can be done as in the proof of Strassen's theorem, Case C. On each box $S_n \times S_m$ let

$$\mu^1_{nm}(C) = \frac{\mu((C \cap S_n) \times S_m)}{\mu(S_n \times S_m)}, \quad \mu^2_{nm}(C) = \frac{\mu(S_n \times (C \cap S_m))}{\mu(S_n \times S_m)}$$

be the marginal distributions of the conditional distribution of μ on $S_n \times S_m$. Define

$$\mu' = \sum_{n,m} \mu(S_n \times S_m)\ \mu^1_{nm} \times \mu^2_{nm}.$$

In this construction, locally on each small box $S_n \times S_m$, measure μ is replaced by the product measure with the same marginals. Let us compute the marginals of μ'. Given a set $C \subseteq S$,

$$\begin{aligned} \mu'(C \times S) &= \sum_{n,m} \mu(S_n \times S_m)\ \mu^1_{nm}(C) \times \mu^2_{nm}(S) \\ &= \sum_{n,m} \mu((C \cap S_n) \times S_m) = \sum_n \mu((C \cap S_n) \times S) = \sum_n \mathbb{P}(C \cap S_n) = \mathbb{P}(C). \end{aligned}$$

Similarly, $\mu'(S \times C) = \mathbb{Q}(C)$, so μ' has the same marginals as μ, $\mu' \in M(\mathbb{P}, \mathbb{Q})$. It should be obvious that transportation cost integral does not change much by replacing μ with μ'. One can visualize this by looking at what happens locally on each small box $S_n \times S_m$. Let (X_n, Y_m) be a random pair with distribution μ restricted to $S_n \times S_m$ so that

$$\mathbb{E}d(X_n, Y_m) = \frac{1}{\mu(S_n \times S_m)} \int_{S_n \times S_m} d(x,y)d\mu(x,y).$$

Let Y'_m be an independent copy of Y_m, also independent of X_n, i.e. the joint distribution of (X_n, Y'_m) is $\mu^1_{nm} \times \mu^2_{nm}$ and

$$\mathbb{E}d(X_n, Y'_m) = \int_{S_n \times S_m} d(x,y)d(\mu^1_{nm} \times \mu^2_{nm})(x,y).$$

Then

$$\int d(x,y)d\mu(x,y) = \sum_{n,m} \mu(S_n \times S_m)\mathbb{E}d(X_n, Y_m),$$

$$\int d(x,y)d\mu'(x,y) = \sum_{n,m} \mu(S_n \times S_m)\mathbb{E}d(X_n, Y'_m).$$

Finally, $d(Y_m, Y'_m) \le \mathrm{diam}(S_m) \le \varepsilon$ and these two integrals differ by at most ε. Therefore,

$$\int d(x,y)d\mu'(x,y) \le W(\mathbb{P}, \mathbb{Q}) + 2\varepsilon.$$

Similarly, we can define

$$\nu' = \sum_{n,m} \nu(S_n \times S_m)\ \nu^1_{nm} \times \nu^2_{nm}$$

such that

$$\int d(x,y)d\nu'(x,y) \le W(\mathbb{Q}, \mathbb{T}) + 2\varepsilon.$$

We will now show that this special simple form of the distributions $\mu'(x,y), \nu'(y,z)$ ensures that the conditional distributions of x and z given y are well defined. Let $\mathbb{Q}_m$ be the restriction of $\mathbb{Q}$ to S_m,

$$\mathbb{Q}_m(C) = \mathbb{Q}(C \cap S_m) = \sum_n \mu(S_n \times S_m)\ \mu^2_{nm}(C).$$

Obviously, if $\mathbb{Q}_m(C) = 0$ then $\mu^2_{nm}(C) = 0$ for all n, which means that μ^2_{nm} are absolutely continuous with respect to $\mathbb{Q}_m$ and the Radon-Nikodym derivatives

$$f_{nm}(y) = \frac{d\mu^2_{nm}}{d\mathbb{Q}_m}(y) \ \text{ exist and } \ \sum_n \mu(S_n \times S_m) f_{nm}(y) = 1 \ \text{ a.s. for } y \in S_m.$$

Let us define a conditional distribution of x given y by

$$\mu'(A|y) = \sum_{n,m} \mu(S_n \times S_m) f_{nm}(y) \mu^1_{nm}(A).$$

Notice that for any $A \in \mathcal{B}$, $\mu'(A|y)$ is measurable in y and $\mu'(A|y)$ is a probability distribution on $\mathcal{B}$, $\mathbb{Q}$-a.s. over y because

$$\mu'(S|y) = \sum_{n,m} \mu(S_n \times S_m) f_{nm}(y) = 1 \ a.s.$$

Let us check that for Borel sets $A, B \in \mathcal{B}$,

$$\mu'(A \times B) = \int_B \mu'(A|y)d\mathbb{Q}(y).$$

Indeed, since $f_{nm}(y) = 0$ for $y \notin S_m$,

$$\begin{aligned}
\int_B \mu'(A|y)d\mathbb{Q}(y) &= \sum_{n,m} \mu(S_n \times S_m)\mu^1_{nm}(A) \int_B f_{nm}(y) d\mathbb{Q}(y) \\
&= \sum_{n,m} \mu(S_n \times S_m)\mu^1_{nm}(A) \int_B f_{nm}(y) d\mathbb{Q}_m(y) \\
&= \sum_{n,m} \mu(S_n \times S_m)\mu^1_{nm}(A)\mu^2_{nm}(B) = \mu'(A \times B).
\end{aligned}$$

Conditional distribution $\nu'(\cdot|y)$ can be defined similarly.

$\square$

Next lemma shows that on a separable metric space any law with the "first moment", i.e. $\mathbb{P} \in \mathcal{P}_1(S)$, can be approximated in metrics W and γ by laws concentrated on finite sets.

Lemma 42 *If (S,d) is separable and $\mathbb{P} \in \mathcal{P}_1(S)$ then there exists a sequence of laws $\mathbb{P}_n$ such that $\mathbb{P}_n(F_n) = 1$ for some finite sets F_n and $W(\mathbb{P}_n, \mathbb{P}), \gamma(\mathbb{P}_n, \mathbb{P}) \to 0$.*

Proof. For each $n \geq 1$, let $(S_{nj})_{j\geq 1}$ be a partition of S such that $\mathrm{diam}(S_{nj}) \leq 1/n$. Take a point $x_{nj} \in S_{nj}$ in each set S_{nj} and for $k \geq 1$ define a function

$$f_{nk}(x) = \begin{cases} x_{nj}, & \text{if } x \in S_{nj} \text{ for } j \leq k, \\ x_{n1}, & \text{if } x \in S_{nj} \text{ for } j > k. \end{cases}$$

We have,

$$\int d(x, f_{nk}(x))d\mathbb{P}(x) = \sum_{j\geq 1} \int_{S_{nj}} d(x, f_{nk}(x))d\mathbb{P}(x) \leq \frac{1}{n}\sum_{j\leq k} \mathbb{P}(S_{nj}) + \int_{S\setminus(S_{n1}\cup\cdots\cup S_{nk})} d(x, x_{n1})d\mathbb{P}(x) \leq \frac{2}{n}$$

for k large enough because $\mathbb{P} \in \mathcal{P}_1(S)$, i.e. $\int d(x, x_{n1})d\mathbb{P}(x) < \infty$, and the set $S \setminus (S_{n1} \cup \cdots \cup S_{nk}) \downarrow \emptyset$.

Let μ_n be the image on $S \times S$ of the measure $\mathbb{P}$ under the map $x \to (f_{nk}(x), x)$ so that $\mu_n \in M(\mathbb{P}_n, \mathbb{P})$ for some $\mathbb{P}_n$ concentrated on the set of points $\{x_{n1}, \ldots, x_{nk}\}$. Finally,

$$W(\mathbb{P}_n, \mathbb{P}) \leq \int d(x,y)d\mu_n(x,y) = \int d(f_{nk}(x), x)d\mathbb{P}(x) \leq \frac{2}{n}.$$

Since $\gamma(\mathbb{P}_n, \mathbb{P}) \leq W(\mathbb{P}_n, \mathbb{P})$, this finishes the proof.

□

We are finally ready to extend Theorem 49 to separable metric spaces.

Theorem 50 *(Kantorovich-Rubinstein) If (S,d) is a separable metric space then for any two distributions $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_1(S)$ we have $W(\mathbb{P}, \mathbb{Q}) = \gamma(\mathbb{P}, \mathbb{Q})$.*

Proof. By previous lemma, we can approximate $\mathbb{P}$ and $\mathbb{Q}$ by $\mathbb{P}_n$ and $\mathbb{Q}_n$ concentrated on finite (hence, compact) sets. By Theorem 49, $W(\mathbb{P}_n, \mathbb{Q}_n) = \gamma(\mathbb{P}_n, \mathbb{Q}_n)$. Finally, since both W, γ are metrics,

$$\begin{aligned} W(\mathbb{P}, \mathbb{Q}) &\leq W(\mathbb{P}, \mathbb{P}_n) + W(\mathbb{P}_n, \mathbb{Q}_n) + W(\mathbb{Q}_n, \mathbb{Q}) \\ &= W(\mathbb{P}, \mathbb{P}_n) + \gamma(\mathbb{P}_n, \mathbb{Q}_n) + W(\mathbb{Q}_n, \mathbb{Q}) \\ &\leq W(\mathbb{P}, \mathbb{P}_n) + W(\mathbb{Q}_n, \mathbb{Q}) + \gamma(\mathbb{P}_n, \mathbb{P}) + \gamma(\mathbb{Q}_n, \mathbb{Q}) + \gamma(\mathbb{P}, \mathbb{Q}). \end{aligned}$$

Letting $n \to \infty$ proves that $W(\mathbb{P}, \mathbb{Q}) \leq \gamma(\mathbb{P}, \mathbb{Q})$.

□

Wasserstein's distance $W_p(\mathbb{P}, \mathbb{Q})$. Given $p \geq 1$, let us define the Wasserstein distance $W_p(\mathbb{P}, \mathbb{Q})$ on $\mathcal{P}_p(\mathbb{R}^n) = \{\mathbb{P} : \int |x|^p d\mathbb{P}(x) < \infty\}$ corresponding to the cost function $d(x,y) = |x-y|^p$ by

$$\begin{aligned} W_p(\mathbb{P}, \mathbb{Q})^p &:= \inf\Big\{\int |x-y|^p d\mu(x,y) : \mu \in M(\mathbb{P}, \mathbb{Q})\Big\} \\ &= \sup\Big\{\int f d\mathbb{P} + \int g d\mathbb{Q} : f(x) + g(y) < |x-y|^p\Big\}. \end{aligned} \tag{20.0.2}$$

Even though for $p > 1$ the function $d(x,y)$ is not a metric, equality in (20.0.2) for compactly supported measures $\mathbb{P}$ and $\mathbb{Q}$ follows from the proof of Theorem 49, which does not require that d is a metric. Then one can easily extend (20.0.2) to the entire space $\mathbb{R}^n$. Moreover, W_p is a metric on $\mathcal{P}_p(\mathbb{R}^n)$ which can be shown the same way as in Lemma 41. Namely, given nearly optimal $\mu \in M(\mathbb{P}, \mathbb{Q})$ and $\nu \in M(\mathbb{Q}, \mathbb{T})$ we can construct $(X,Y,Z) \sim M(\mathbb{P}, \mathbb{Q}, \mathbb{T})$ such that $(X,Y) \sim \mu$ and $(Y,Z) \sim \nu$ and, therefore,

$$W_p(\mathbb{P}, \mathbb{T}) \leq (\mathbb{E}|X-Z|^p)^{\frac{1}{p}} \leq (\mathbb{E}|X-Y|^p)^{\frac{1}{p}} + (\mathbb{E}|Y-Z|^p)^{\frac{1}{p}} \leq (W_p^p(\mathbb{P}, \mathbb{Q}) + \varepsilon)^{\frac{1}{p}} + (W_p^p(\mathbb{Q}, \mathbb{T}) + \varepsilon)^{\frac{1}{p}}.$$

Let $\varepsilon \downarrow 0$.

□